1 Introduction

Semantics in the Webster dictionary is defined as *meaning or relationship of meanings of a sign or a set of signs* [4]. From an Information Systems perspective, semantics of information can be defined as *the meaning and use of information* [1]. The Semantic Web is defined as an extension of the current Web in which information is given a well-defined meaning, better enabling people and computers to work together [2]. The Semantic Web is a vision: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but also for automation, integration and reuse of data across various applications. The goal of the Semantic Web initiative is as broad as that of the Web: to create a universal medium for the exchange of data. It is envisaged to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific and cultural data.

These descriptions, though espousing the same vision and goal, however give rise to multiple interpretations, reflecting the perspectives of various fields of computer science and informatics, which can be enumerated as follows:

- Researchers in the database and information systems communities have developed conceptual, logical and physical data and process models to capture semantics of information and processes. These communities have focussed on efficient and scalable storage, indexing and querying of data on one hand, and efficient choreography and orchestration of workflows and services on the other.
- Researchers in the knowledge representation community have developed expressive knowledge representation schemes and theories to capture semantics of information and processes in a declarative manner. This community has focussed on implementation of reasoners and inference mechanisms for checking validity and satisfiability of knowledge specifications.
- Researchers in the information retrieval community have developed thesauri and taxonomies to capture semantics of information. These thesauri and taxonomies have been used to guide search and browsing of documents in document collections. Statistical approaches to capture latent semantics by computing co-occurrence frequencies of terms in a corpus have also been developed.
- Researchers in the machine learning and natural language processing committees have focussed on semantic annotations of data and documents with respect to a well defined set of categories and concepts. Recently there have been efforts related to learning ontologies or taxonomies of concepts.

- Semantics of information has been used to support efficient distributed computing related to location of relevant resources. Researchers in the peer-to-peer communities have proposed approaches that use semantic annotations and localized mappings to locate resources and perform data integration.
- Researchers in the agent systems communities have used ontologies to represent both the semantics of the messages exchanged between agents and the protocols followed by a community of agents for performing a set of tasks.
- Recently researchers in the Web Services communities have proposed process models and ontologies to capture the semantics of services and to a limited extent the semantics of computations to enable reuse and interoperability of applications.

Information technology and the Web has become ubiquitous in our day-to-day lives. Web-based approaches have become the default mode for implementing business processes and operations. Different communities and market verticals such as telecommunication, manufacturing, healthcare and the life sciences are using information and Web-based systems in significant ways to streamline their processes and gain competitive advantage in the market. Approaches that put machine understandable data on the Web are becoming more prevalent.

We discuss next the motivations behind and the vision and goals of the Semantic Web. A framework for characterization of the technologies behind the Semantic Web is presented. A use case for the Semantic Web from the domain of healthcare and the life sciences is presented. This will be the underlying scenario used to motivate the various Semantic Web technologies that will be discussed in this book. Finally, the organization of this book into various units and chapters will be presented.

1.1 Motivation: Why Semantic Web?

With Internet connectivity everywhere and an overabundance of available information the infrastructure for communication is in place. Still, information and services are distributed, often hard to find and hard to integrate. This results in a higher cost to find relevant information and get value from it. Several aspects are fueling the Semantic Web effort:

- Increased cost pressure and competition require businesses to reduce costs by interconnecting workflows and business processes and simplifying the effort of data and service sharing.
- Portal implementations within organizations and e-Government in almost every developed country are aiming to unify the access to government information and services.
- Organizations are attempting to increase automation and interoperability by publishing machine-interoperable data on the Web. Scalability and interopera-

bility across multiple information systems within and across communities has become an urgent priority.

- Scientific progress requires a stronger collaboration and intra- and inter-community information sharing. Various efforts in these areas are aiming to enable data and services sharing: examples are, among many more, the Gene Ontology and Bio Ontology Working Groups for genomics data, the CME project of the Southern California Earthquake Center for seismological data and services among seismologists, and GEON, a geosciences project aiming to interlink and share multi-disciplinary data-sets.
- E-business efforts have focussed on creating business-specific vocabularies for information and process interoperation across enterprises within and across industry boundaries. Some examples of these vocabularies are BPMI, XML-HR and CIM/DMTF.
- There have been efforts in the healthcare and life sciences areas to create specialized vocabularies for different domains. The Unified Medical Language Thesaurus (UMLS®) is a collection of various biomedical vocabularies used for capturing diagnoses (SNOMED), billing information (ICD9) and search queries (MeSH). Life sciences researchers have developed specialized vocabularies such as Gene Ontology for capturing information about genomic structure, function and processes; BioPax for capturing information about Biological Pathways; and MAGE-ML for representing micro-array data in a standardized format.
- The Web is increasingly used as a collaboration forum, using blogs, wikis and other tools for sharing information and tasks such as collaborative metadata annotation and ontology building.

However, this has led to the design and implementation of a multitude of data and metadata schemes along with specialized workflows and processes within and across communities, giving rise to the problem of information overload and the “Tower of Babel” phenomenon. The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently. This is one of the main goals and motivation behind the Semantic Web vision.

1.2 A Framework for Semantic Web

There is a widespread misperception that the Semantic Web is primarily a rehash of existing AI and database work focussed on encoding KR formalisms in markup languages such as RDF(S), DAML+OIL or OWL. We seek to dispel this notion by presenting the broad dimensions of this emerging Semantic Web and the multi-disciplinary technological underpinnings. In fact, we argue that it is absolutely critical to be able to seek, leverage and synergize contributions from a wide variety of technologies and sub-fields of computer science. A framework for presenting the Semantic Web viewed from different perspectives is:

- **Information Aspects of the Semantic Web**:
  Semantic Web Content. This refers to the myriad forms of data that will be presented on the Semantic Web along with the metadata descriptions embedded in the data. This is best exemplified by the equation:
  Semantic Web Content = Data + Metadata.
  Data. This includes structured (e.g., relational) data, semi-structured (e.g., RDF, XML) data and unstructured (e.g., raw text multimedia) data consisting of metadata descriptions embedded in the data.
  Metadata and Annotations. This refers to the various types of domain- or application- specific metadata descriptions that will be used to annotate data on the Semantic Web. Annotation is fundamental to the creation of the Semantic Web.
  Ontologies and Schemas. This refers to the underlying vocabulary and semantics of the metadata annotations. Collections of domain-specific concepts may be used to create domain- and application-specific views on the underlying content. Schemas are a special case of metadata that are structured and may contain semantic information. In the cases where the metadata is explicit (e.g., database schemas, XML schemas), these metadata may be mapped to other related metadata or ontological concepts.

Table 1.1. Framework for the Semantic Web

	Information Aspects			**Process Aspects**
	Content	**Metadata and Annotations**	**Ontologies and Schemas**	
DB and CM Systems[a]	X	X		
KR Systems[b]		X	X	
Machine Learning		X	X	
Statistical Clustering		X	X	
Information Retrieval	X	X		
NLP[c]	X	X	X	
Distributed Computing				X
SOA	X	X	X	X
Agents	X	X	X	X

Table 1.1. Framework for the Semantic Web

	Information Aspects			**Process Aspects**
	Content	**Metadata and Annotations**	**Ontologies and Schemas**	
P2P[d]		X	X	X
Grid Computing				X

a. DB and CM systems refer to database and content management systems respectively

b. KR Systems refer to Knowledge Representation Systems

c. NLP refers to Natural Language Processing Technologies

d. P2P refers to Peer-to-Peer Computing Infrastructures

- **Computational Aspects of the Semantic Web**:

  Computing Infrastructures. This refers to various computing infrastructures that support the communication between computational entities such as agents, P2P, Web Services and the different styles of communication.

  Direct Communication. Direct communication enables computational entities to send synchronous and asynchronous messages to each other through direct communication channels.

  Mediated Communication. Mediated communication supports a mediator that acts as intermediary for two or more computational entities to communicate.

  Service Description. This refers to the definition of the interfaces of computational entities that will communicate with each other. The description needs to ensure interoperability of applications at the semantic level through the definition of service behavior as well as dynamic service composition and invocation. Service-level agreements governing the scalability and performance are part of service descriptions.

1.3 Use Case: Translational Medicine Clinical Vignette

The field of translational medicine may be defined as in [3]: (a) validation of theories emerging from preclinical experimentation on disease-affected human subjects and; (b) refinement of biological principles that underpin human disease heterogeneity and polymorphisms by using information obtained from preliminary human experimentation. This is a new emerging field which straddles the health ecosystem, consisting of diverse market sectors such as healthcare delivery, drug discovery and life sciences. In this section, we present a high-level description of a clinical vignette which has ramifications across the health ecosystem. This clinical vignette will be used throughout the book to motivate solutions based on Semantic Web technologies and discuss details related to the same. A detailed description of the clinical vignette, identifying high-level functional requirements will be presented in the next chapter.

The use case begins when a patient enters a doctor's office complaining of fainting-like symptoms and pain in the chest. The doctor performs a clinical examination of the patient that reveals abnormal heart sounds. On further discussion with the patient, the doctor learns that the patient has a family history of sudden death, with the patient's father dying at the age of 40. His two younger brothers are apparently normal. The doctor then decides to order an ultrasound based on his clinical examination of the patient. The echocardiogram reveals hypertrophic cardiomyopathy. This could lead to the sequencing of various genomes such as MYH7 and MYBPC3 which could result in the doctor's ordering various types of therapies, drugs and monitoring protocols.

This clinical use case will be used to motivate the need for various aspects of semantics-related technologies:

- Semantics-rich data and information models that can capture phenotypic information related to the physical exam, structured image reports and family history on one the hand; and on the other hand genotypic information such as the results of molecular diagnostic tests such as mutations, expression levels, and so on.
- Semantic mappings that would enable integration of data and information across clinical and genomic data sources. Some of these mappings may be complex and require execution of rule-based specifications.
- The ability to specify semantic rules for implementing clinical decision support that could suggest appropriate molecular diagnostic tests based on the phenotypic characteristics of the patient and propose therapies based on the patient's genotype.
- The ability to semi-automate knowledge acquisition of genotypic and phenotypic associations and other knowledge that could inform decision support and provide a substrate for information integration. Statistical and Machine Learning techniques are specially relevant in this context.
- The ability to manage change in knowledge due to the rapid rate of knowledge discovery in the healthcare and life sciences. The role of semantic inferences based on expressive ontologies and information models will be crucial in this context.
- The ability to represent processes such as therapeutic protocols and biological pathways and possible ways of combining them. Semantic Web process models are likely to be useful in this context.
- The ability to orchestrate clinical workflows across multiple clinical and genomic contexts.

1.4 Scope and Organization

In this book, we will discuss the state of the art related to the use and deployment of Semantic Web specifications. These discussions will be presented in the context of a framework presented in Section 1.2. We focus on a very pragmatic view of the Semantic Web, viz., that of a way of standardizing data, information, knowledge and process specifications to achieve enhanced information and process interoperability, decision support and knowledge change management. We will focus on how these semantic technologies and specifications can work in the context of a use case discussed in the previous section. We will not discuss issues related to the "strong AI" such as higher order logics, modeling of consciousness and cognitive abilities of human beings.

The organization of the book is as follows:

- Chapter 2 presents a detailed discussion of the clinical vignette discussed earlier. High-level use case description and functional requirements and architectural assumptions are presented. These will be used to motivate the various Semantic Web technologies discussed in this book.
- Chapter 3 discusses various examples of different types of "Semantic Web" content spanning across structured, semi-structured and unstructured data.
- Chapter 4 discusses various metadata frameworks based on W3C recommendations such as XML, RDF and OWL. Issues related to the data models of markup specifications such as XML, RDF and OWL; and associated query languages such as XML Schema and SPARQL are discussed with examples drawn from a solution approach to the clinical use case .
- Chapter 5 discusses the broad question of *What is an Ontology?* and various artifacts created by different communities such as thesauri, schemas and classification schemes will be discussed. Specifications for representation of schemas and ontologies on the Web such as XML Schema, RDF Schema and OWL are discussed with examples drawn from a solution approach to the clinical use case.
- Chapter 6 discusses issues related to ontology authoring, bootstrapping and management. Tools for ontology authoring, merging, versioning and integration are discussed follow by a discussion of ontology versioning and change management issues that arise in the context of the clinical use case.
- Applications enabled by the use of metadata descriptions and ontologies on the Semantic Web, are presented in Chapter 7. Tools for metadata annotation and are discussed followed by a discussion on approaches for ontology-based information integration.
- A discussion of communication models such agents, P2P, client/server and their relationship to Web Services and the Semantic Web is presented in Chapter 8.
- A discussion of the current standards and the state of the art in Web Services is presented in Chapter 9. The impact of semantic inadequacies on web services based on current web standards will also be discussed.

- Dynamic Services Composition, a key enabler of functionality reuse on the Semantic Web, is discussed in Chapter 10
- Chapter 11 discusses the idea of Semantic Web Services and how the inadequacies of current web standards can be addressed.
- Chapter 12 enumerates standards in the area of Semantic Web and Semantic Web Services.
- Chapter 13 presents a solution based on Semantic Web technologies for the use case requirements presented in Chapter 2.

Finally, the list of references and the list of index entries follow.